Issues in Information Systems
Volume 17, Issue I, pp. 99-106, 2016

DO SOCIAL NETWORKING SITES PRIVACY POLICIES DIFFER? A LINGUISTIC ANALYSIS OF THE TEN MOST POPULAR SOCIAL NETWORKING SITES

Alan R. Peslak, Penn State University, arp14@psu.edu

ABSTRACT

Social networking sites such as Facebook and LinkedIn are used by 65% of the US population and nearly 2 billion people worldwide. A significant amount of personal information is shared on these sites. As a result the privacy policies that these sites follow is critical. Our study reviews the privacy policies of the top ten social networking sites and reviews them from a linguistic perspective using LIWC (Linguistic and Word Count) analysis. A detailed analysis by site and comparisons across sites is performed. We found significant differences in several linguistic measures including Analytic, Clout, Tone, and use of pronouns. The Implications of these differences and other key findings are discussed.

Keywords: Social Networking, Privacy Policy, Linguistic Analysis, LIWC, Facebook, Content Analysis, Qualitative Analysis

INTRODUCTION

Since the invention of the Internet, information has been easily shared by billions of users across the world on an ongoing basis. Initially scientific, academic, news and reference information were the primary data that was shared across the Internet. Gradually as the Internet expanded, more individuals and personal information began to be shared. And thus began the concept of social networking sites. The most cited journal article on social networking (Ellison, 1997) defines social networking as "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system." Since that time many other sites have been created and successfully found users and markets. It is now reported that 65% of the US population (Perrin, 2015) and 2 billion people worldwide (Statistica, 2016) regularly use some form of social networking sites. But privacy issues remain with social networking sites. Facebook and others constantly update their privacy policies to keep up with user concerns but concerns continue. As an example, Liu, Y., Gummadi, K. P., Krishnamurthy, B., & Mislove, A. (2011) found that Facebook privacy setting do not meet user expectations 63% of the time. Obviously, privacy concerns remain and are a fertile topic for research and analysis. Our study reviews the privacy policies of the ten largest social networking sites from a unique perspective, linguistic analysis. We use the most researched and accepted analysis tool LIWC (Linguistic and Word Count) software, in order to glean insight and perspective into these privacy policies. The goal is to add to the existing research and determine significant insights into the current state of privacy within social networking.

LITERATURE REVIEW

Due to its ubiquitous presence, there is extensive literature on the use of social networking. Social networking is a broad category of information sharing hosted sites on the Internet between individuals and organizations. According to the most cited journal article on Social Networking history (Ellison, 2007) "Since their introduction, social network sites (SNSs) such as MySpace, Facebook, Cyworld, and Bebo have attracted millions of users, many of whom have integrated these sites into their daily practices. As of this writing, there are hundreds of SNSs, with various technological affordances, supporting a wide range of interests and practices. While their key technological features are fairly consistent, the cultures that emerge around SNSs are varied. Most sites support the maintenance of pre-existing social networks, but others help strangers connect based on shared interests, political views, or activities. Some sites cater to diverse audiences, while others attract people based on common language or shared racial, sexual, religious, or nationality-based identities. Sites also vary in the extent to which they incorporate new

information and communication tools, such as mobile connectivity, blogging, and photo/video-sharing. Scholars from disparate fields have examined SNSs in order to understand the practices, implications, culture, and meaning of the sites, as well as users' engagement with them." The prevalence and universal nature of social networking is rivaled only by the use of search engines on the Internet. It is important to understand rules and regulations with regard to usage of data provided to these sites. The rules related to how these sites use your information is provided in their privacy policy.

Many prior studies have been performed on privacy policies of Internet sites. Jensen and Potts (2004) examined privacy policies as a decision making rule. Miyazaki and Krishnamurthy (2002) studied the relationship between privacy policies and consumer perception.

Also, many studies have been performed on the specific issue of social networking privacy in the literature. Dwyer, Hiltz and Passerini (2007) studied concepts of trust and privacy in Facebook and Myspace. They found concerns with trust in both sites and in other members of the sites. Livingstone (2008) studied the levels of risks that teenagers take with regard to privacy in their usage of social networking sites. Barnes (2006) developed a comprehensive overview of the privacy issues associated with social networking usage. Gerlach, Widjaja,, & Buxmann (2015) found that privacy policies are changing beyond just ecommerce concerns. Almeur, Lawani, and Dalkir (2016) conducted a experiment on traditional versus personalized privacy policies and show that allowing personalization and management in privacy policies affects user trust. Qi and Edgar-Nevill (2011) studied the need and mechanisms for improved data collection and privacy control from web site data gathering.

The usage of linguistic analysis on privacy policies has not been found to have been performed in the past. But the usage of linguistic analysis and specifically the use of LIWC (Linguistic and Word Count) software for research purposes has been extensive. Robinson, Navea, and Ickes (2013) used LIWC analysis of students written self-introductions to successfully predict course performance. Back, Kufner, and Egloff (2011) analyzed 9/11 communications using LIWC. Bell, McCarthy, and McNamara (2012) used LIWC to investigate gender differences in linguistic styles. Sexton and Helmreich (2000) studied airline cockpit communications via LIWC to determine errors and performance. Cordova, Cunningham, Carlson, and Andrkowski (2001) used LIWC to analyze how individuals adjusted to having breast cancer. There are many more examples of the use of LIWC in the literature. The use of LIWC has been well established and accepted in peer-reviewed journals.

LIWC (Linguistic and Word Count) software (Pennebaker, Booth, Boyd, and Francis, 2015) is the most researched and popular linguistic analysis tool. "The way that the **L**inguistic **I**nquiry and **W**ord **C**ount (LIWC) program works is fairly simple. Basically, it reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech. Because LIWC was developed by researchers with interests in social, clinical, health, and cognitive psychology, the language categories were created to capture people's social and psychological states. …. LIWC reads written or transcribed verbal texts which have been stored in a digital, computer-readable form (such as text files). The text analysis module then compares each word in the text against a user-defined dictionary. As described below, the dictionary identifies which words are associated with which psychologically-relevant categories.

After the processing module has read and accounted for all words in a given text, it calculates the percentage of total words that match each of the dictionary categories. For example, if LIWC analyzed a single speech that was 2,000 words and compared them to the built-in LIWC2015 dictionary, it might find that there were 150 pronouns and 84 positive emotion words used. It would convert these numbers to percentages, 7.5% pronouns and 4.2% positive emotion words."(Pennebaker Conglomerates, 2015).

METHODOLOGY

Our linguistic analysis will take two forms. First, we will review the results of each social networking privacy policy. Next we will review the top 10 social networking sites privacy policies to see if there are significant linguistic and content differences between their policies. We will discuss important results and conclusions.

RQ1 What information can be determined from a linguistic review of the major social networking sites privacy policies.

RQ2 Are there significant linguistic and content differences between the major social networking sites?

The first step was to determine the most popular social networking sites. Fortunately, ebizmab.com publishes the most popular social networking sites based on Global and US traffic. The data was used for March 2016 and the top 10 sites were used for our analysis. They are in order: Facebook, Twitter, LinkedIn, Pinterest, Google Plus+, Tumblr, Instagram, VK, Flicker, Vine. (Ebizmba, 2016). All websites were then searched and their respective privacy policies downloaded and saved. (Note that Google Plus+ had a unique policy but also referenced the overall Google policy so this was also included, downloaded and saved. Flickr is owned by Yahoo and governed by their policy which was downloaded and saved).

The next step was to process each privacy policy via LIWC and perform the analyses to address the research questions. This follows in the results section.

RESULTS

RQ1 What information can be determined from a linguistic review of the major social networking sites privacy policies.

LIWC software results produce 93 unique measures from each of its linguistic analyses. These measures range from parts of speech to emotional categories to word counts. For the most part these are expressed by a percentage of total words mapping to the dictionary category of each measure. The exceptions are several relating to word counts as well as calculated emotional measures. It would be prohibitive to analyze each policy and explain each measure so only a select group of each of the measures will be explained for each policy in RQ1. The authors studied results and extracted these major results.

The actual results of our analyses are presented in Table 2. But in order to understand these results we have provided Pennebaker's measures for selected media as well as a basic explanation of each metric in table 1.

For comparison purposes, the selected measures and the average score/percentage of words in each category for major media are presented in table 1 along with a basic explanation. The results for each of the top ten sites for those measures as calculated by the LIWC software are shown in table 2.

Table 1. LIWC Analytic Measures for Selected Media
(Pennebaker, J., Boyd, R., Jordan, K., and Blackburn, K. (2015))

	Blogs	Expressive writing	Novels	Natural Speech	NY Times	Twitter	Explanation
Analytic	49.89	44.88	70.33	18.43	92.57	61.94	Logical versus informal
Clout	47.87	37.02	75.37	52.27	68.17	63.02	Confident versus humble
Authentic	60.93	76.01	21.56	61.32	24.84	50.39	Honest versus guarded
Tone	54.5	38.6	37.06	79.29	43.51	72.24	Upbeat versus hostile
Words Per Sentence	18.4	18.42	16.13		21.94	12.1	Reading complexity
Words greater than 6 letters	14.38	13.62	16.3	10.42	23.58	15.31	Educated
Dictionary Words	85.79	91.93	84.52	91.6	74.62	82.6	Nontechnical

Function	53.1	58.27	54.51	56.86	42.39	46.08	Style versus content
Pronoun	16.2	16.2	18.03	15.15	20.92	7.41	Personal and informal
Positive emotions	3.66	2.57	2.67	5.31	2.32	5.48	Happy
Negative emotions	2.06	2.12	2.08	1.19	1.45	2.14	Sad or angry
Cognitive processes	11.58	12.52	9.84	12.27	7.52	9.96	requiring thought

With this basic background in what each metric represents, we can now examine the results of our LIWC analysis of the top ten social networking sites. These are presented in table 2. There are many similarities and differences between these results and what is presented in other forms of communication shown in table 1. These differences that we found are discussed in the results section.

Table 2 LIWC Results for Selected Measures for Top Ten Social Networking Sites (Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015)).

Filename	Fbook	Flickr	Google	Instagrm	Linkedin	Pintrst	Tumblr	Twitter	Vine	VK
Analytic	54.12	76.09	67.89	68.55	67.69	53.22	62.88	67.34	65.87	99.00
Clout	99.00	97.86	98.25	97.78	99.00	99.00	99.00	99.00	99.00	69.15
Authentic	24.92	27.32	26.94	37.65	39.13	35.84	19.66	31.02	34.15	17.91
Tone	85.68	72.99	68.95	74.38	82.79	76.94	84.26	68.79	69.35	46.77
WPS	20.59	21.83	22.69	26.43	23.82	20.55	26.16	27.04	26.52	21.96
Sixltr	28.22	29.07	29.47	29.26	31.19	26.97	27.69	30.76	30.47	26.87
Dic	86.75	80.82	79.94	82.55	82.41	85.31	85.31	84.38	83.33	72.20
Function	51.84	44.61	46.15	49.09	48.08	50.89	51.40	48.26	48.84	42.50
Pronoun	16.90	12.93	14.59	14.74	15.36	18.10	16.55	14.58	15.45	5.80
Posemo	3.83	2.74	2.75	2.84	3.39	2.77	3.73	2.37	2.37	1.41
Negemo	0.40	0.21	0.46	0.23	0.20	0.00	0.42	0.09	0.06	0.28
Cogproc	15.92	14.34	16.82	17.65	16.38	20.42	16.62	15.22	16.67	10.00

RQ2 Are there significant linguistic and content differences between the major social networking sites?

For the selected measures explored in Research Question 1, we next performed a One-Sample Kolmogorov-Smirnov non-parametric test to determine if there were significant differences in the LIWC measures. The results are shown in table 3.

Table 3. LIWC Selected Results One-Sample Kolmogorov-Smirnov Test

One-Sample Kolmogorov-Smirnov Test	N	Normal Parameters a,b		Test Statistic	Asymp. Sig. (2-tailed)
		Mean	Std. Deviation		
Analytic	10	3405.9	1819.854	0.280476	.025c
Clout	10	1	.000d		
Authentic	10	68.265	12.77344	0.2911	.016c
Tone	10	95.704	9.343666	0.487914	.000c

Words Per Sentence	10	29.454	7.353781	0.138453	.200c,e
Words greater than 6 letters	10	73.09	11.23597	0.250971	.074c
Dictionary	10	23.759	2.576201	0.22433	.167c
Words	10	28.997	1.539488	0.130669	.200c,e
Function	10	82.3	4.125188	0.210637	.200c,e
Pronoun	10	48.166	3.009512	0.188601	.200c,e
Positive emotions	10	14.5	3.37954	0.309443	.007c
Negative emotions	10	2.82	0.713396	0.188817	.200c,e
Cognitive processes	10	0.235	0.157357	0.152813	.200c,e
Analytic	10	16.004	2.651608	0.187364	.200c,e

One sample K-S test for differences

DISCUSSION

Table 1 shows the LIWC results for our selected variables that Pennebaker et al. found when reviewing different forms of communications. They used a variety of formal and informal communications from thousands of samples to come up with average scores and percentages for each LIWC measure. We will use this table as one of the comparisons for our analyzed social networking privacy policies.

The results of our LIWC analysis are shown in tables 2 and 3. Table 2 shows actual results for each of the measures for the top ten social networking sites. Table 3 provides a one sample K-S one-parametric test to determine if there are statistically significant differences between the measures for each site. Overall, we found significant differences in Analytic, Clout, Tone, and pronouns at $p < .10$. All except Tone were significant at $p < .05$. We will next analyze each privacy policies and each metric.

The LIWC analysis of the privacy policies first sees a high level of guardedness in all of the top ten privacy policies (henceforth referred to as policies). The authenticity (Authentic) measure averages only 29.5. "Higher numbers are associated with a more honest, personal, and disclosing text; lower numbers suggest a more guarded, distanced form of discourse." (Pennebaker, Booth, Boyd, and Francis, 2015). Our results suggest a less personal form of disclosure and more distanced. The most direct comparison is that it is approximately equal to news reporting as evidenced by the NY Times score. It does not have the more honest and personal style of blogs, expressive writing or Twitter. One proposed explanation is that the policies are more factual and do not reflect ambiguous or interpretive content. All the policies have a relatively low level of authenticity, with none over a score of 40 on the 100 point scale. There is no significant difference between the ten policies.

All of the policies also have a similar level of complexity and education level as evidenced by the Words per Sentence and greater than six letter words measures. The 24 words per sentence and 29% over six letter words exceed even the NY Times and suggest a high level of reading level and educational level for understanding. This may be an issue and present challenges for understanding the policies by less educated individuals. There is no significant differences in these measures.

All the policies except VK exhibit an extremely high level of clout or expertise "Clout ------ a high number suggests that the author is speaking from the perspective of high expertise and is confident; low Clout numbers suggest a more tentative, humble, even anxious style." (Pennebaker, Booth, Boyd, and Francis, 2015). The exception of VK is most probably due to limitations on privacy policies placed on European Union members since VK is the largest European social network. The Eu has a higher level of privacy protection for individuals and is most probably reflected here. Once again, there is no significant difference.

The Analytic measure shows interesting results. "Analytical thinking ------ a high number reflects formal, logical, and hierarchical thinking; lower numbers reflect more informal, personal, here --- and --- now, and narrative

thinking." The most analytic policy is exhibited by VK, with a near perfect formal score. The rest of the policies show varying degrees of formality from a neutral formality of 54 for Facebook to a much higher Flickr formality at 76. This may in part be due to the near exclusive form of posts as pictures for Flickr versus text for many others. Differences in the Analytic measure are significant at $p < .016$.

The Emotional tone of the policies is statistically different at $p < .074$. "Emotional tone ------ a high number is associated with a more positive, upbeat style; a low number reveals greater anxiety, sadness, or hostility. A number around 50 suggests either a lack of emotionality or different levels of ambivalence." Most have a moderate level of positive tone except VK which is ambivalent at 46. Of the rest Facebook is highest and most positive at 86 and twitter is lowest at 69. In general, though most social networks try to keep an upbeat emotional tone.

Function words are non-content words in a document such as articles, prepositions, and pronouns. Overall there is no significant difference in function (style) words versus content words. VK had the lowest function word score (42) and thus the highest content. This a level similar to the NY Times. Facebook had the highest function words at 52. This is the level of Blog writing.

Dic or percent of words found in the dictionary is suggested to represent technical complexity of the writing. A low level suggest less common, more difficult and technical writing. The NY Times score is 75. All of our top ten social networking privacy policies exceeded 80 in score and are less technical. The exception to this however was again VK which had a score of 72, suggesting higher technical and writing complexity. There was no statistical difference found though in the dictionary metric.

The use of pronouns suggests a more personal and informal style. The usage of pronouns in most of the top ten policies are less than the NY Times, novels, expressive writing, natural speech or blogs with the average at 14.5. The higher sites are Facebook, Pinterest, and Tumblr which have more pronouns and thus more informal style than the others. There is a statistically significant difference in this metric. VK is very low on this scale at 5.8.

The positive and negative emotion metrics show the percentage of words that display either positive or negative emotions. For the most part, the inclusion of emotions is positive emotion. With an average of 2.8% positive emotion words, the policies exceed that of novels, expressive writing, and the NY Times. What is particularly noteworthy though is the near absence of any negative emotion words. With only .24%, none of the external categories come close to this almost nil level. Also the proportion of positive to negative words in more than 10 to 1. Only natural speech comes close to this proportion at 5 to 1. Twitter is only a little more than 2 to 1. The policies are not significantly different in these measures.

The final measure we discuss is cognitive process. These are words that suggest thinking or requiring some mental activity. This measure is higher than any form of general communications: blogs, expressive writing, novels, natural speech, NY Times, or Twitter. The policies are not significantly different in this measure.

CONCLUSION

Overall, this study has demonstrated a series of important results. First it defines, presents and demonstrates an example and interpretation of linguistic analysis using one of the most researched and developed tools, LIWC. Researchers and practitioners can use this manuscript as a source and guide for developing their own linguistic analysis of any communication. Second, the study illustrates the results of privacy policies metrics as they compare to other forms of communications. These policies compare to Novels in Analytic, higher than any in Clout and Cognitive processes, fairly guarded like the NY Times, highly upbeat like Twitter, have NY Times complexity, are about as non-technical and personal as Blogs, and have a very high ratio of positive to negative emotion. Researchers and practitioners can reliably use this comparison for other forms of communications. Finally, the study analyzes the privacy policies and sentiment of the ten largest social networking sites. The results show significant differences in many areas of linguistic analyses. There are significant differences in Analytic, Clout, Tone, and pronoun use. There is no significant difference in Authentic, Words Per Sentence, Over Six letter words, Dictionary percentage, Positive or negative emotions or cognitive processes. Researchers can use these findings to compare to other social network policies or other privacy policies for other type sites to compare and contrast their linguistic

characteristics. Social networking companies can use these findings to improve their overall sentiment if they choose.

REFERENCES

Aïmeur, E., Lawani, O., & Dalkir, K. (2016). When changing the look of privacy policies affects user trust: An experimental study. *Computers in Human Behavior*, *58*, 368-379.

Back, M. D., Küfner, A. C., & Egloff, B. (2011). Automatic or the people? Anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, *22*(6), 837-838.

Barnes, S. B. (2006). A privacy paradox: Social networking in the United States. *First Monday*, *11*(9).

Bell, C. M., McCarthy, P. M., & McNamara, D. S. (2012). Using LIWC and Coh-Metrix to investigate gender differences in linguistic styles. *Applied Natural Language Processing: Identification, Investigation, and Resolution, Information Science Reference, Hershey, PA*, 545-556.

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns flexibility in writing style and physical health. *Psychological science*, *14*(1), 60-65.

Cordova, M. J., Cunningham, L. L., Carlson, C. R., & Andrykowski, M. A. (2001). Social constraints, cognitive processing, and adjustment to breast cancer. *Journal of consulting and clinical psychology*, *69*(4), 706.

Dwyer, C., Hiltz, S., & Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. *AMCIS 2007 proceedings*, 339.

Ebizmba (2015). Top 15 Most Popular Social Networking Sites | March 2016 http://www.ebizmba.com/articles/social-networking-websites

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210-230.

Gerlach, J., Widjaja, T., & Buxmann, P. (2015). Handle with care: How online social network providers' privacy policies impact users' information sharing behavior. *The Journal of Strategic Information Systems*, *24*(1), 33-43.

Jensen, C., & Potts, C. (2004, April). Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 471-478). ACM.

Liu, Y., Gummadi, K. P., Krishnamurthy, B., & Mislove, A. (2011, November). Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference* (pp. 61-70). ACM.

Livingstone, S. (2008). Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression. *New media & society*, *10*(3), 393-411.

Miyazaki, A. D., & Krishnamurthy, S. (2002). Internet seals of approval: Effects on online privacy policies and consumer perceptions. *Journal of Consumer Affairs*, *36*(1), 28-49.

Qi, M., & Edgar-Nevill, D. (2011). Social networking searching and privacy issues. *Information Security Technical Report*, *16*(2), 74-78.

Pennebaker Conglomerates (2015). *LIWC How It Works*. http://liwc.wpengine.com/how-it-works/

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015*). Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).,

Pennebaker, J., Boyd, R., Jordan, K., and Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX. University of Texas at Austin.

Perrin, A. (2015). Social Networking Usage: 2005-2015. Pew Research Center. October 2015. Available at: http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology*, 0261927X13476869.

Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: the links between language, performance, error, and workload. *Journal of Human Performance in Extreme Environments*, *5*(1), 6.

Statista (2016). *Statistics and facts about Social Networks*. http://www.statista.com/topics/1164/social-networks/esence